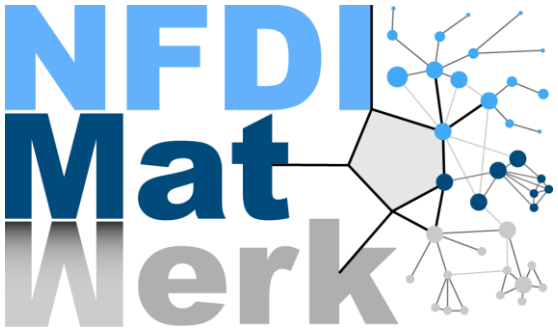
# Exploiting FAIR Data to Enhance Data Analysis

3rd International Network Meeting of EUSMAT
29.06.2022

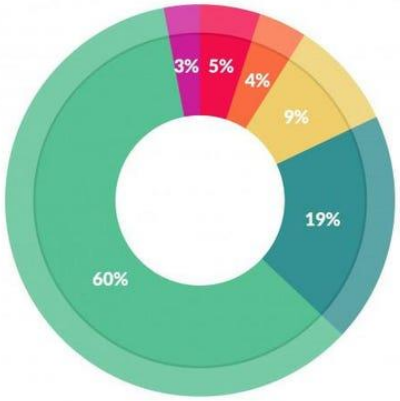
Dr. Marius Politze (RWTH Aachen University)





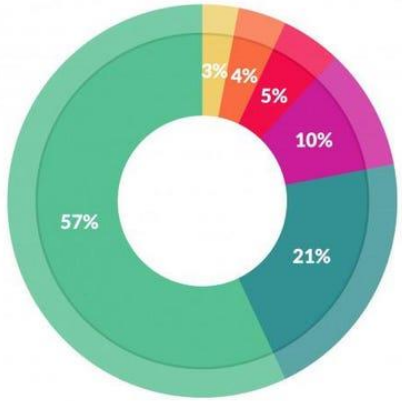

IT Center | RWTH AACHEN UNIVERSITY

# Why Should I care?




What data scientists spend the most time doing

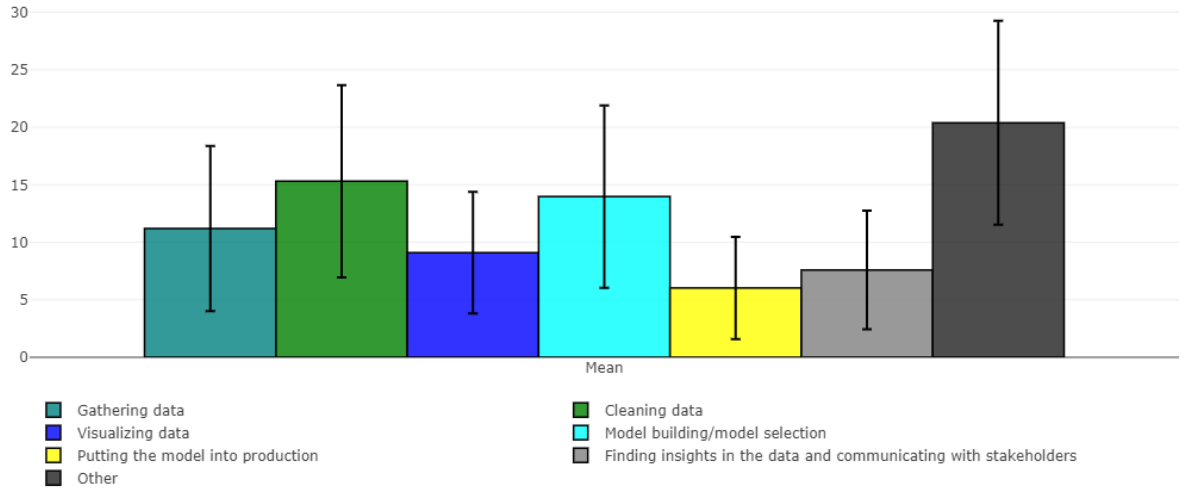
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



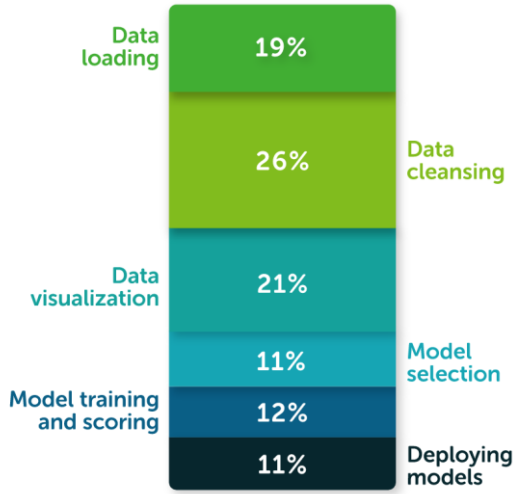

What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Press, G (2016): Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#1b963786f637




Mooney, P (2018): Kaggle Machine Learning & Data Science Survey, https://www.kaggle.com/code/paultimothymooney/2018-kaggle-machine-learning-data-science-survey/notebook




Anaconda Inc. (2020): 2020 State of Data Science, https://www.anaconda.com/state-of-data-science-2020

# Why Should I care?

Data preparation and cleansing takes valuable time away from real data science work and has a negative impact on overall job satisfaction.

Anaconda Inc. (2020): 2020 State of Data Science,
https://www.anaconda.com/state-of-data-science-2020

… and (most of the time) this is because most data collection are a mess!

Marius Politze, today

# FAIR Principles[1]

- Findable:
  Persistent identifiers; indexed and searchable metadata
- Accessible:
  Retrievable using standard protocols; “tombstones”
- Interoperable:
  Vocabularies for data and metadata; qualified references to other (meta)data
- Reusable:
  Licenses; provenance; meets community standards


FAIR guiding principles for data resources, Sangya Pundir, CC-BY-SA-4.0

[1] Wilkinson, M. D. et. al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. doi:10.1038/sdata.2016.18

# Recap FAIR Principles (1)

**To be Findable:**

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

**To be Accessible:**

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

# Recap FAIR Principles (2)

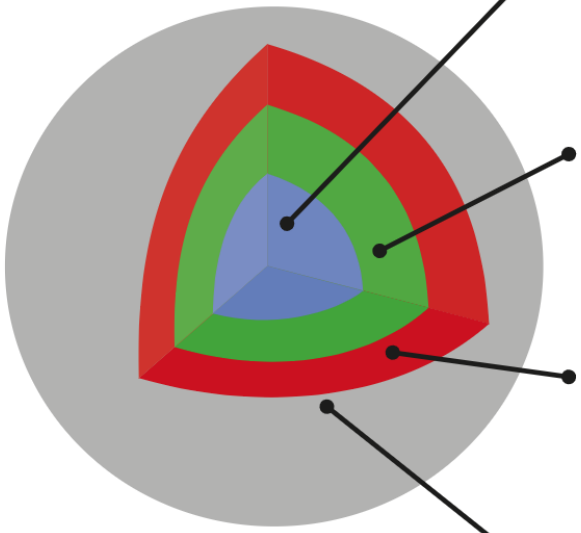
**To be Interoperable:**

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

**To be Re-usable:**

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

# Implementing FAIR Principles: FAIR Digital Objects[1]



DIGITAL OBJECT

**Data, code and other research outputs**

*At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.*

IDENTIFIERS

**Persistent and unique (PIDs)**

*Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).*

STANDARDS & CODE

**Open, documented formats**

*Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.*

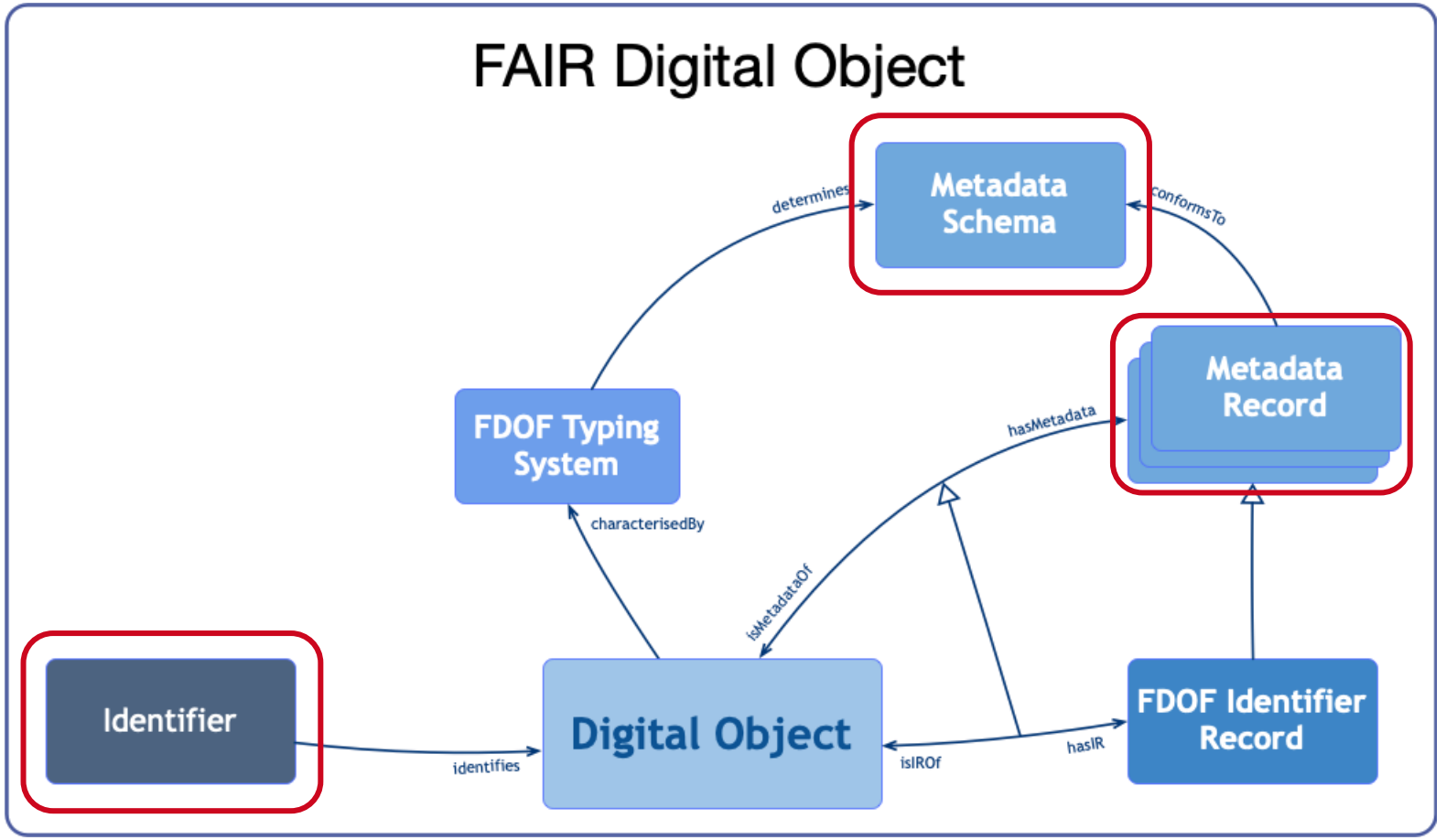
METADATA

**Contextual documentation**

*In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.*

[1] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, doi:10.2777/1524

# FAIR Digital Object – A Structured Set of Links




Bonino da Silva Santos, L. O. (2021): FAIR Digital Object Framework Documentation. https://fairdigitalobjectframework.org/

# Persistent Identifiers - Eternally and Globally Unique

- What is eternity?
  - Whatever the community defines it to be
- What is globally unique?
- Examples
  - **DOI – Published Documents (and Data)**
  - **ORCID – Researchers**
  - ROR – Research Organizations
  - **ePIC/Handle – (Unpublished) Data, real world things**
  - pURL – Mostly digital concepts
  - W3ID – Mostly digital concepts
  - …

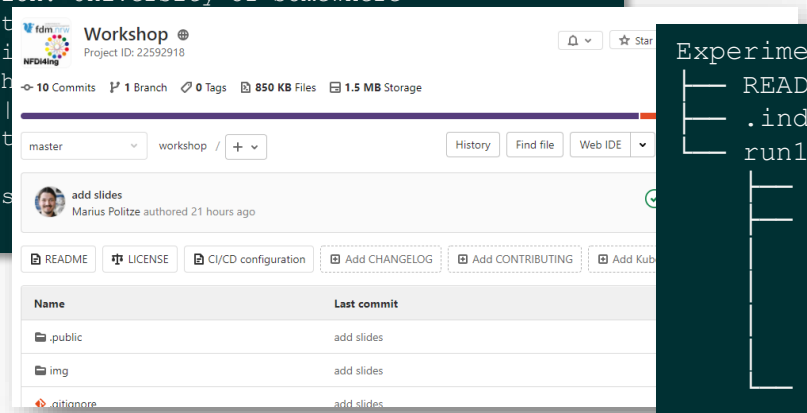
*Alice: How long is forever?*
*White Rabbit: Sometimes, just one second.*
– Alice in Wonderland[1]

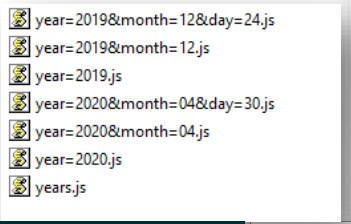
[1] All credits for finding this quote go to York Sure-Vetter (Director of NFDI e.V.)

# (Structured) Metadata Records

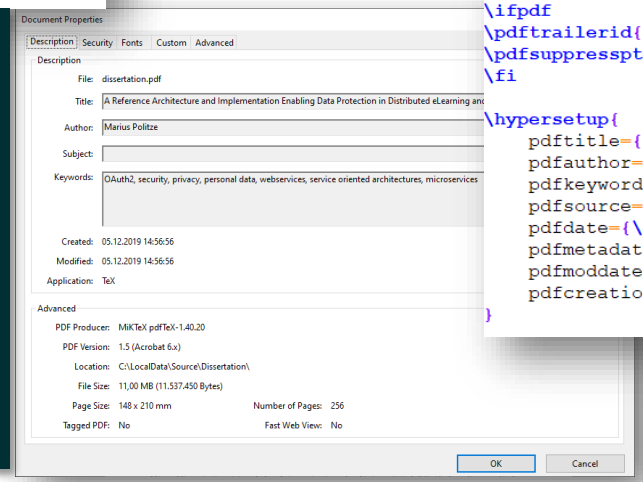
- Text-Files
- Metadata In Files
- Databases
- XML / JSON
- Need for standardization across different storage methods

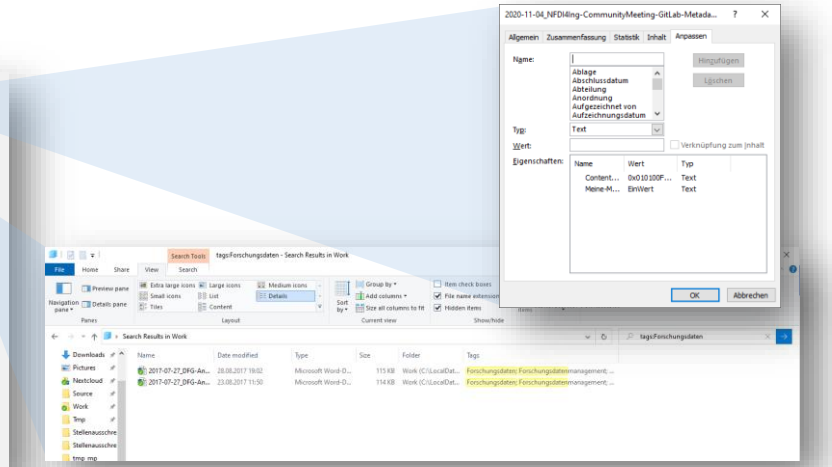
```
---
sensor: sensorA
deviceUnderTest: Prototype C
title:  'This is the title: it contains a colon'
author:
- name: Author One
  affiliation: University of Somewhere
- name: Aut
  affiliati
tags: [noth
abstract: |
  This is t

  It consis
---
```

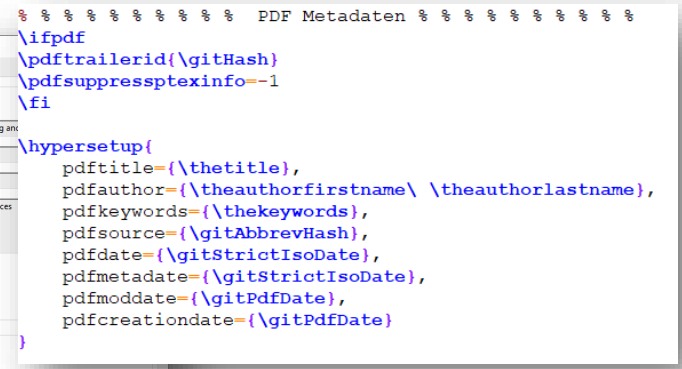



```
Experiment1
├── README.md
├── .index.yml
└── run1
    ├── .index.yml
    ├── sensorA
    │   ├── .index.yml
    │   ├── dataset1.hdf5
    │   ├── dataset1.json
    │   └── dataset2.csv
    └── sensorB
        ├── .index.yml
        └── dataset.abc
```



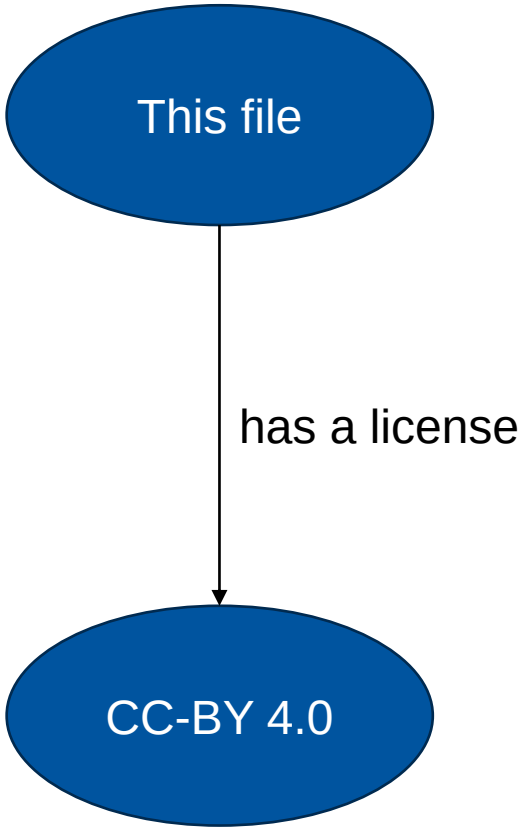





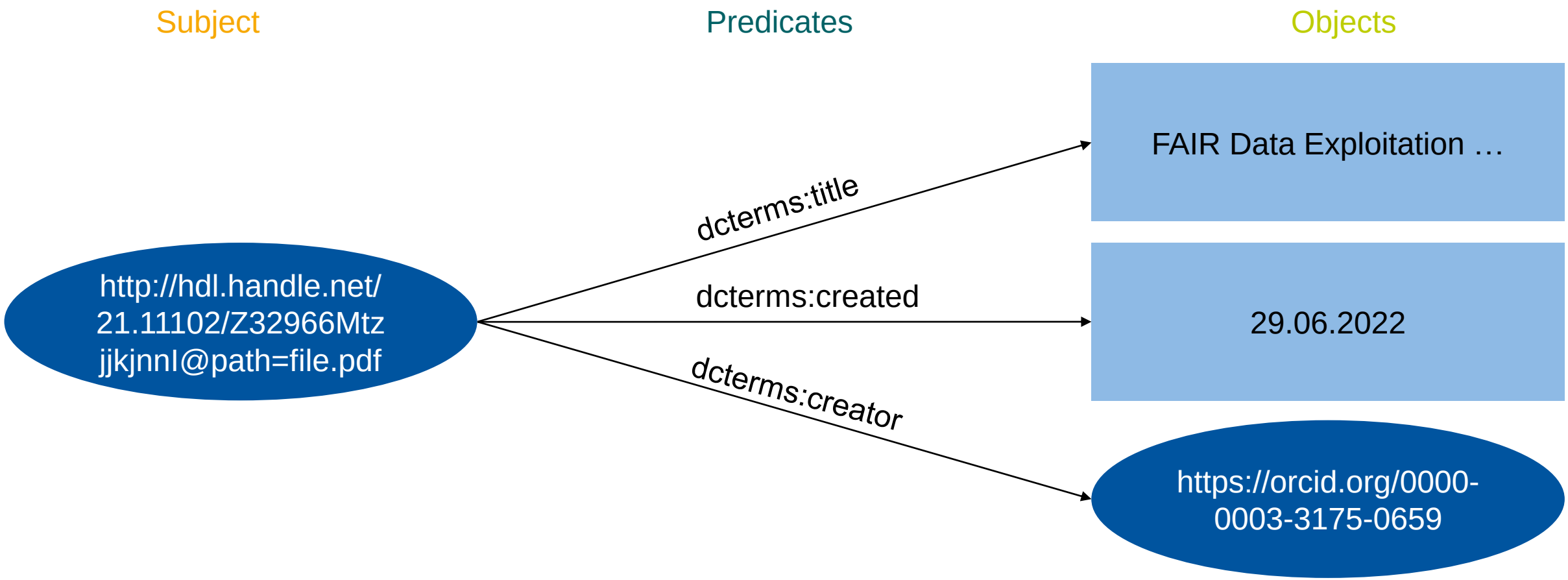
```
% % % % % % % % % %  PDF Metadaten % % % % % % % % % %
\ifpdf
\pdftrailerid{\gitHash}
\pdfsuppressptexinfo=-1
\fi

\hypersetup{
    pdftitle={\thetitle},
    pdfauthor={\theauthorfirstname\ \theauthorlastname},
    pdfkeywords={\thekeywords},
    pdfsource={\gitAbbrevHash},
    pdfdate={\gitStrictIsoDate},
    pdfmetadate={\gitStrictIsoDate},
    pdfmoddate={\gitPdfDate},
    pdfcreationdate={\gitPdfDate}
}
```

# Semi Structured Data on the Web: Linked Data and the Semantic Web

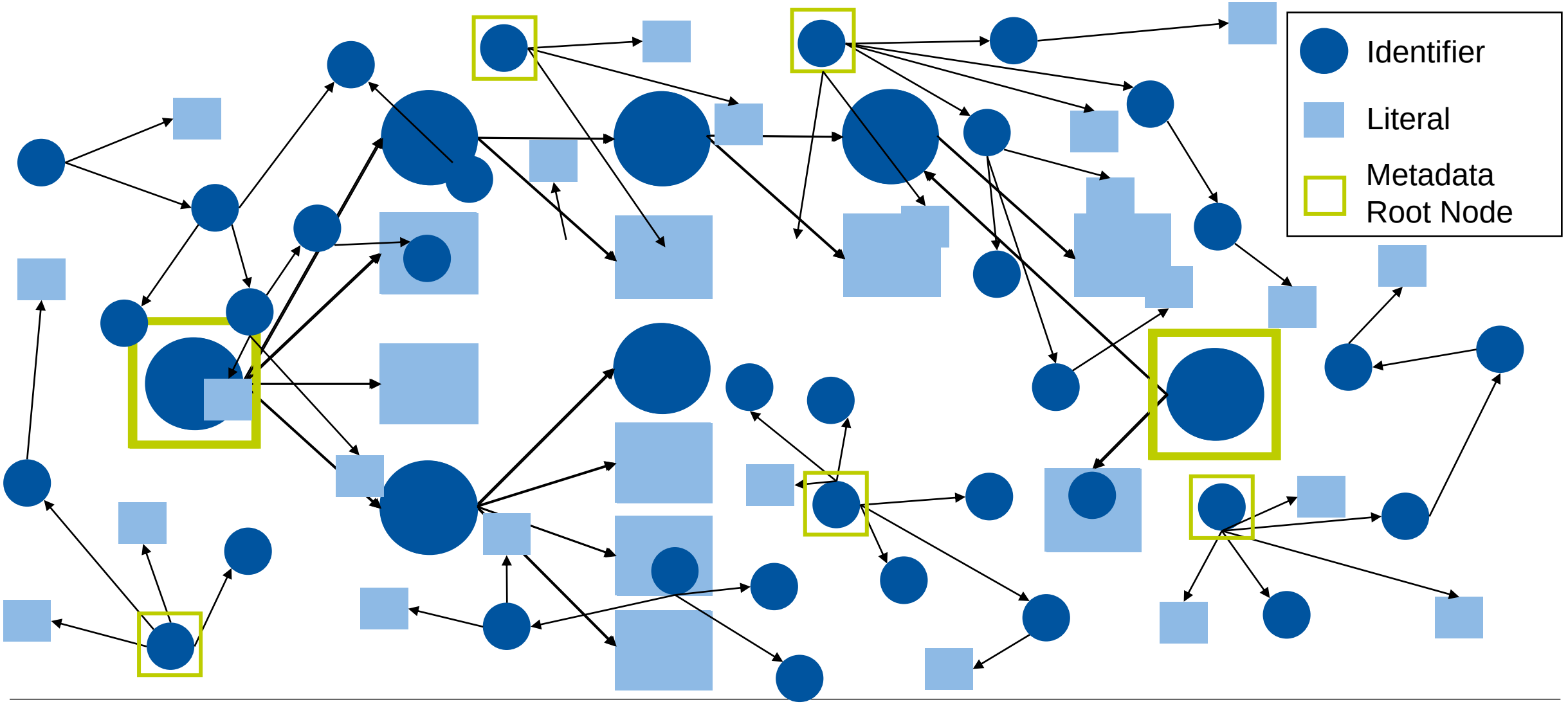
- Structured because the framework is defined → RDF
- RDF defines (meta)data to be saved in triples of “subject predicate object”
  - Think of very simple sentences of the form “A has property of value B
- Unstructured because the terms used can (and should) be freely defined
- Linked Data Terms to uniquely identify concepts
  - Real world  things
  - Digital Objects
  - Relations between these things
  - Abstract properties
- Open Specification allows adding to various formats
  - HTML
  - PDF
  - HDF5
  - …



# RDF Creates a Metadata Graph

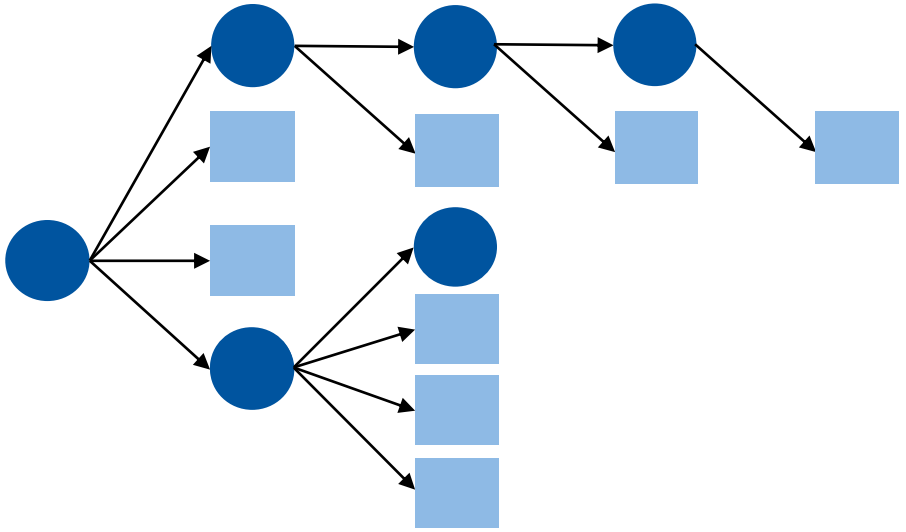


# Metadata as a Knowledge Graph in the Semantic Web



## Bringing Order to the Chaos

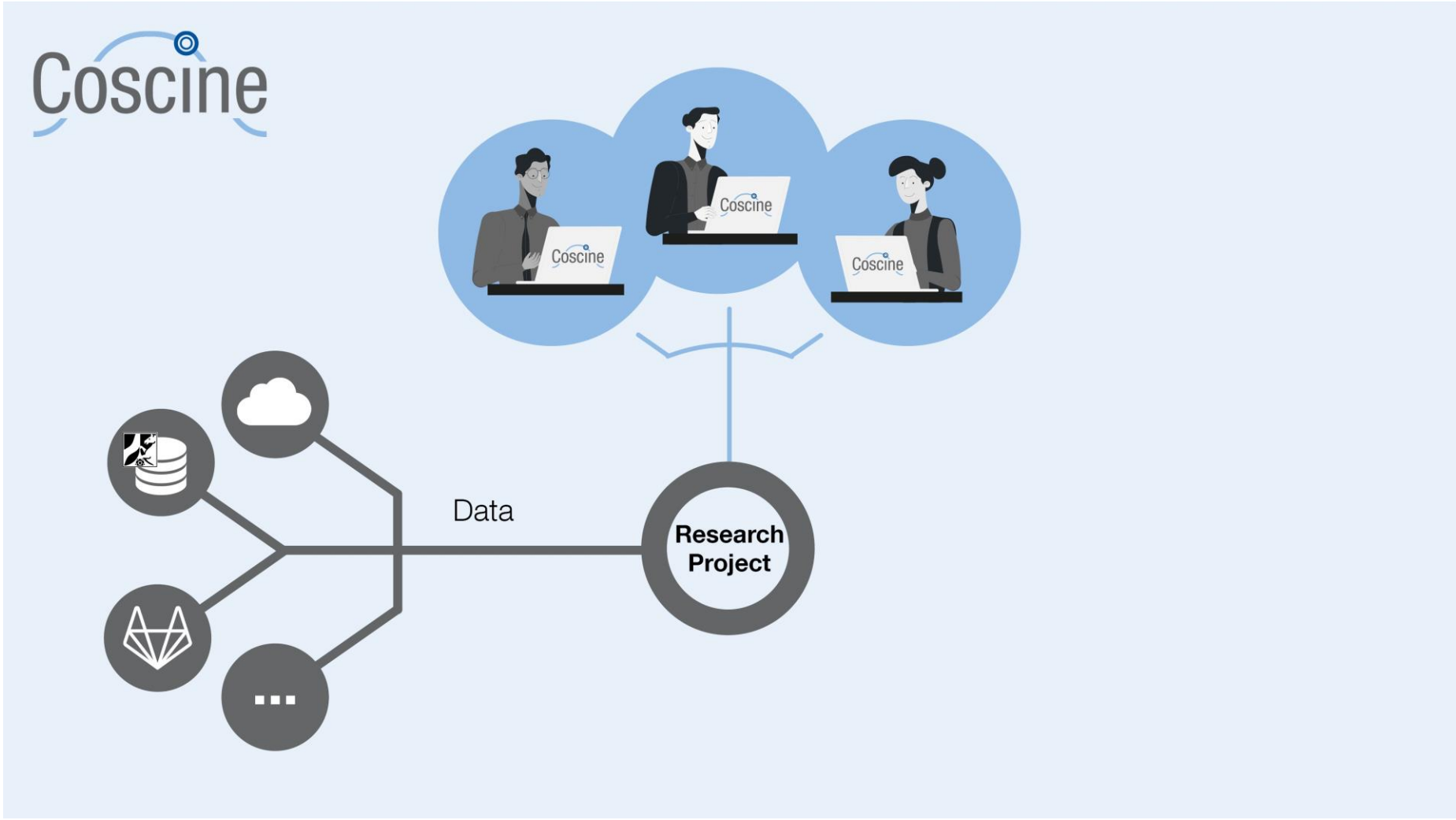
- Metadata Schema
  - “Open World Definition”: Everything is valid if it not explicitly forbidden
  - Set of information to describe things in a given context
  - What do predicates (relations) mean? What are their semantics?
- Metadata Standard
  - A Metadata Schema that is endorsed by some standardization body
- Application Profiles
  - “Cloesed World Definition”:
  - May consist of Information from multiple schemas and standards
  - Define your data structures based on your field of application



# Ok, nice but…

… is there an app for that?

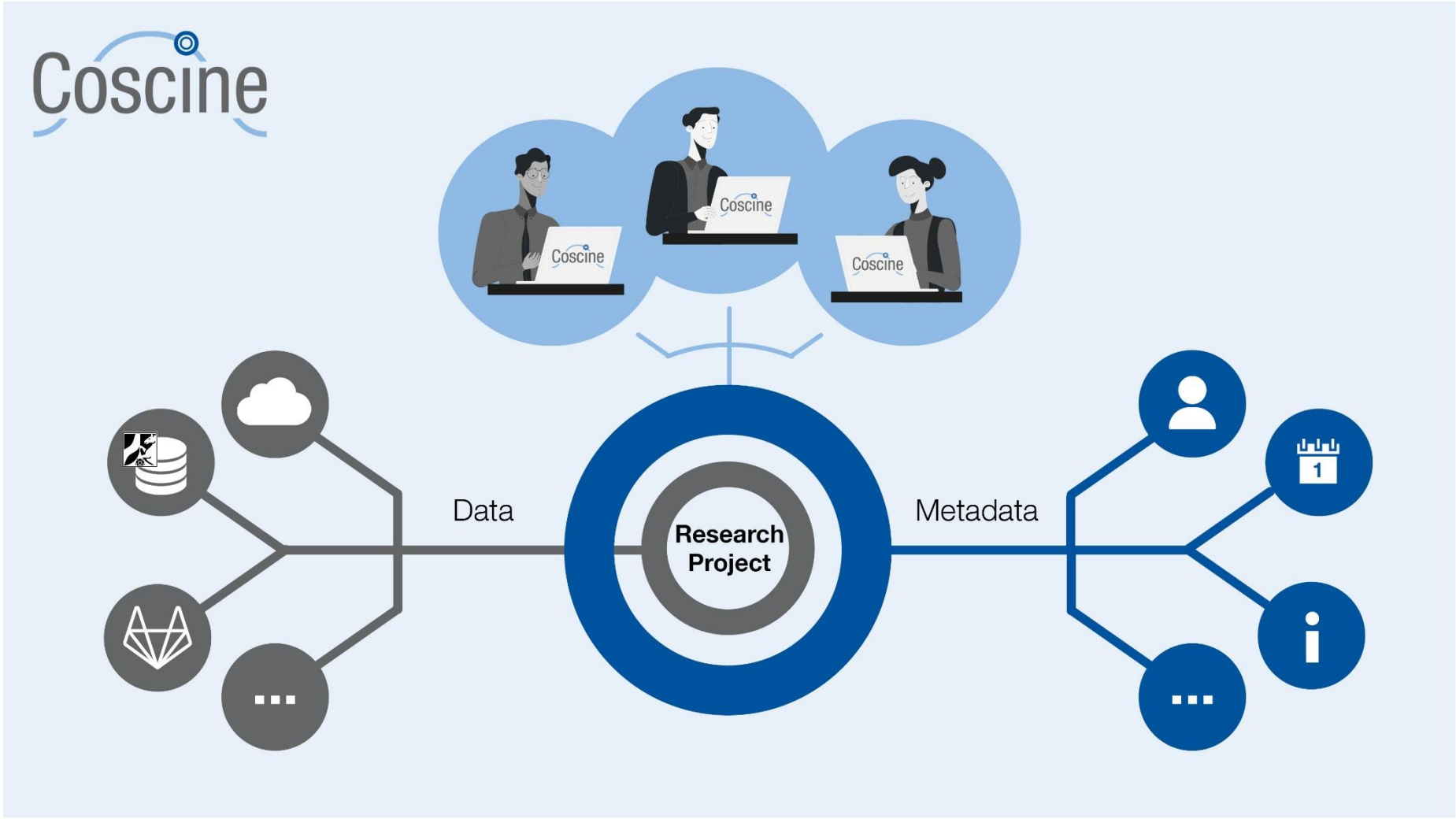
# Coscine - Short Introduction




**Coscine…**

- …maps a project structure
- …integrates different storage systems
- …maintains authorizations at the project level

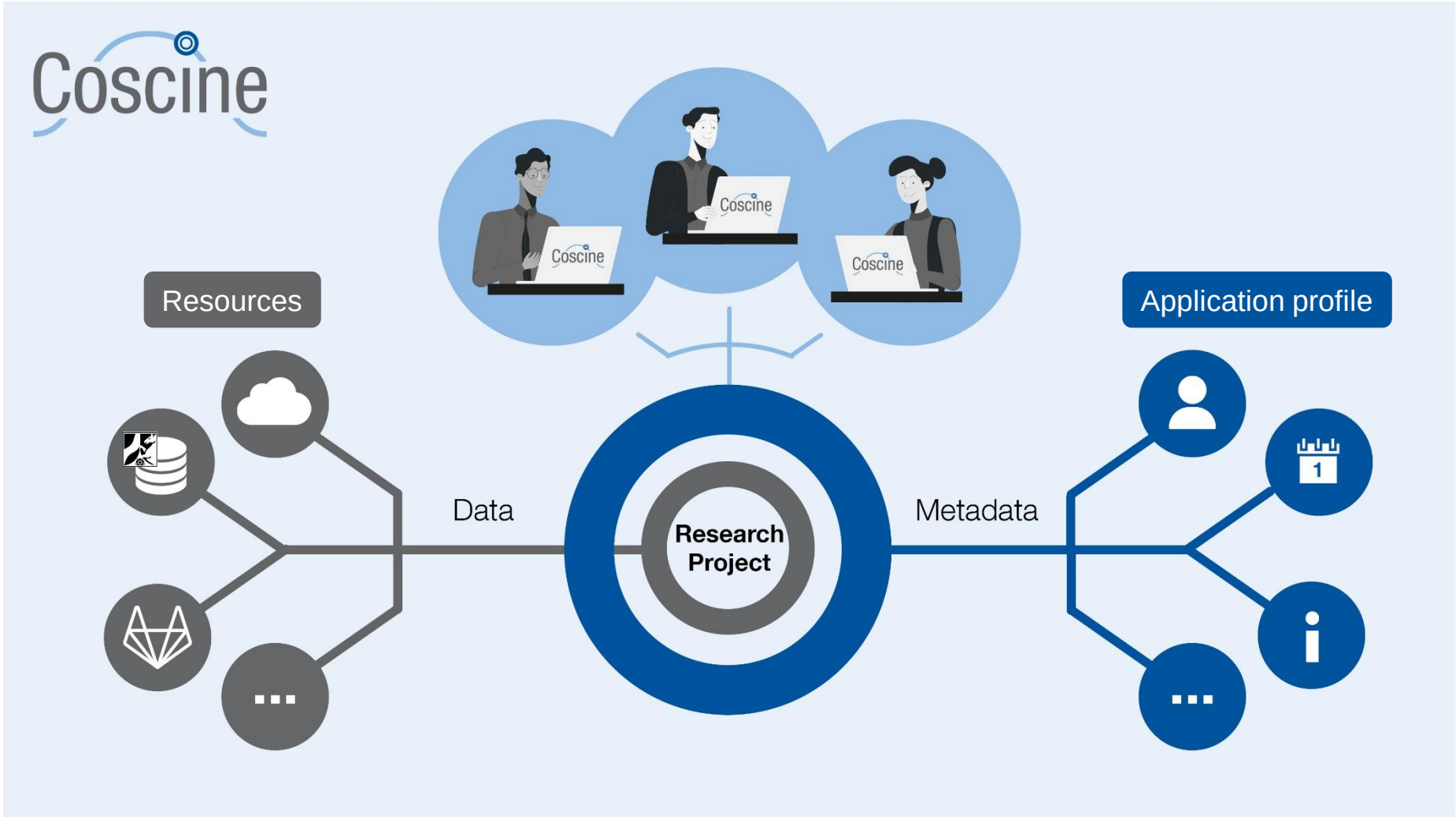
# Coscine - Short Introduction




**Coscine…**

- …maps a project structure
- …integrates different storage systems
- …maintains authorizations at the project level
- …helps to describe all data with structured metadata

# Coscine - Short Introduction

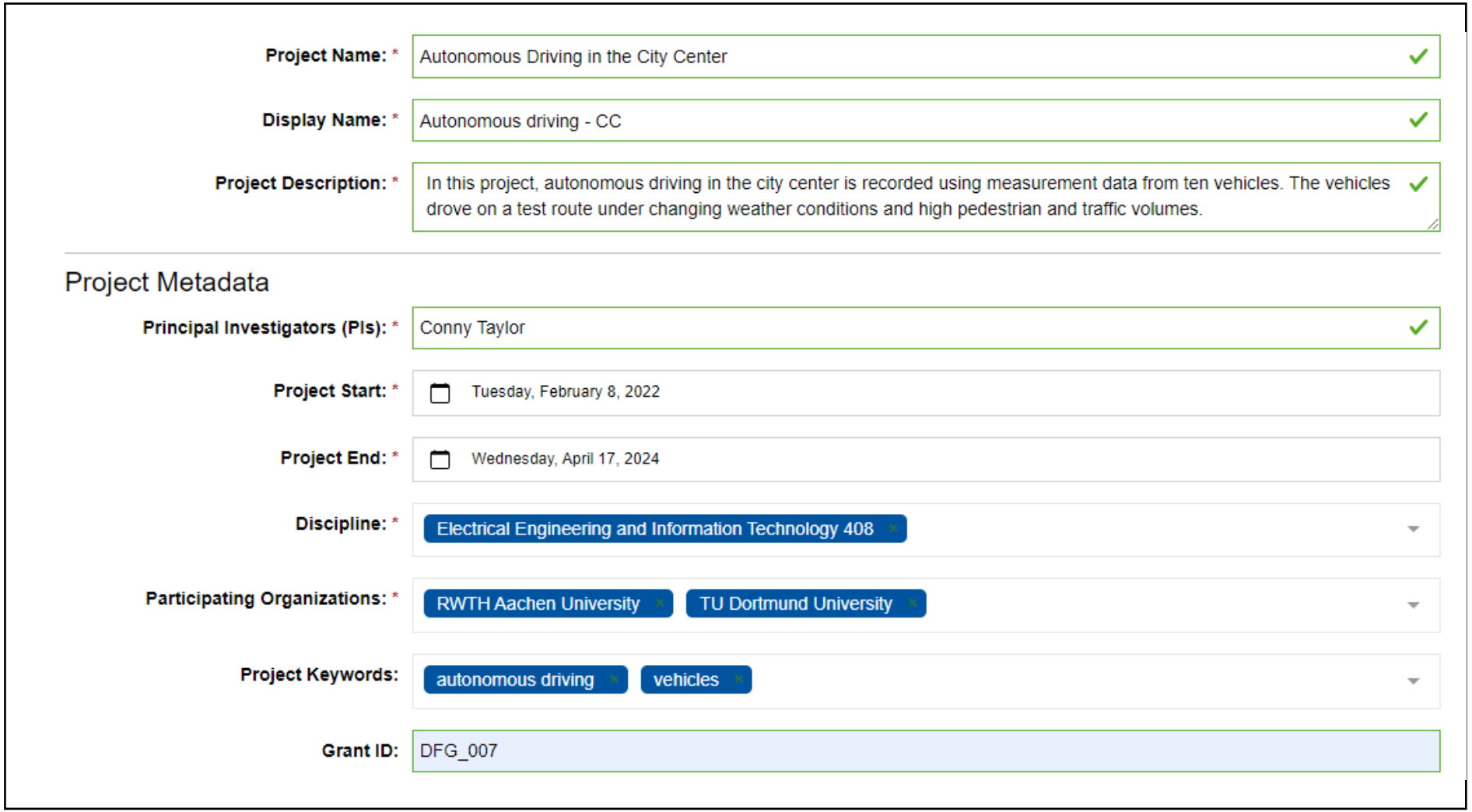



**Coscine…**

- …maps a project structure
- …integrates different storage systems
- …maintains authorizations at the project level
- …helps to describe all data with structured metadata

# Coscine – Metadata Management

## 1. Project Level

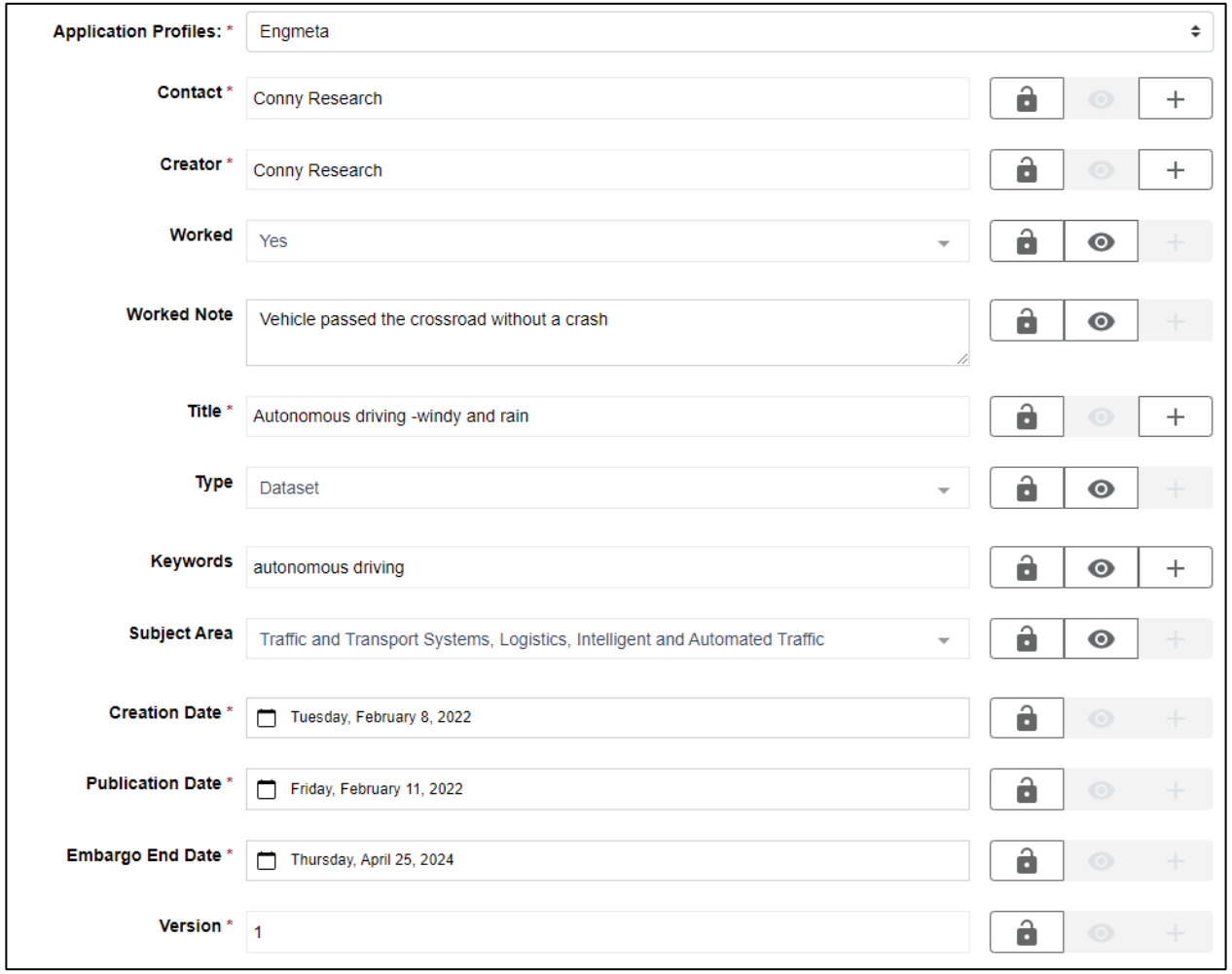


# Coscine – Metadata Management

## 2. Resource Level

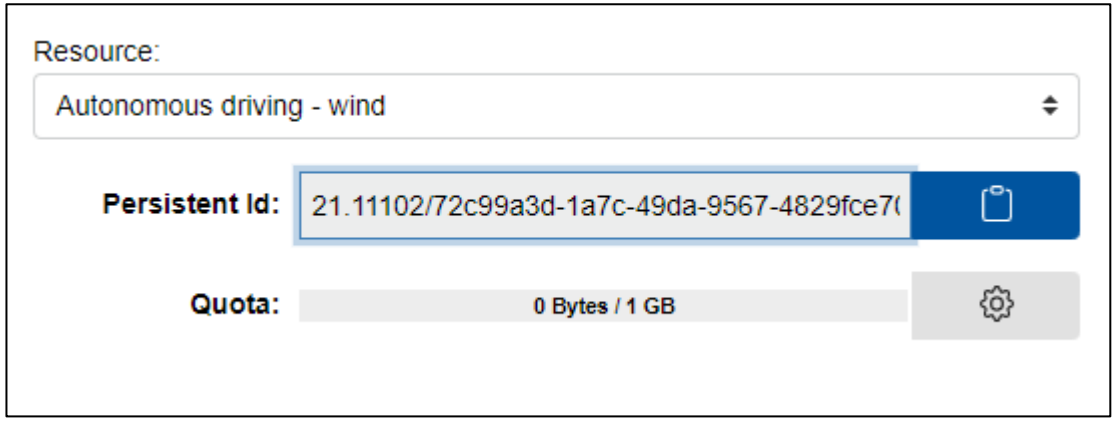
Options:

- **Pre-implemented** application profiles (e.g. EngMeta, see figure)
- Create and request an **individual** application profile



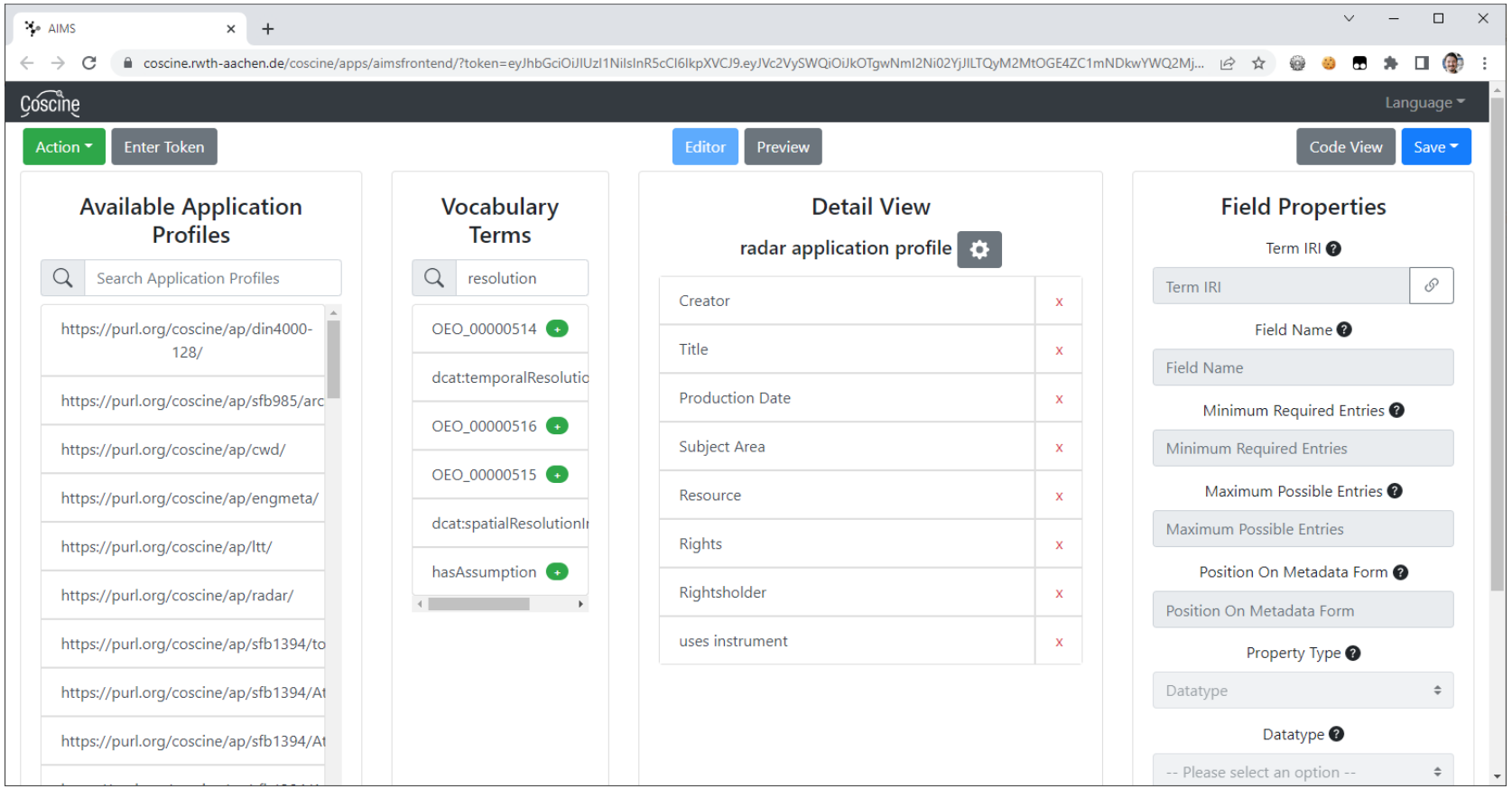
# Coscine – Persistent Identifiers (PIDs)

- Coscine uses PIDs to uniquely reference resources
- Each resource in a project is automatically assigned a PID
- The URL contains the handle-prefix followed by a PID
  - Example: *http://hdl.handle.net/21.11102/7599d318-99f3-4385-ace9-7aeb9cf3bXXX*
- PIDs can be used to link resources and make them accessible to others



# The AIMS Project - Creating Your Own Application Profile

https://coscine.rwth-aachen.de/coscine/apps/aimsfrontend/



# Vielen Dank
# für Ihre Aufmerksamkeit

Dr. Marius Politze
0000-0003-3175-0659

politze@itc.rwth-aachen.de



IT Center | RWTH AACHEN UNIVERSITY